

Linear Regression Assessment & Diagnostics

STAT 245

Sept 24, 2024

Model Diagnostics and Assessment

- Does the model fit data well?
- *Should* we have fit a line -- Is model appropriate for data?
- Are predictors *really* associated with response?

Regression Conditions

L

I

N

E

Our Model

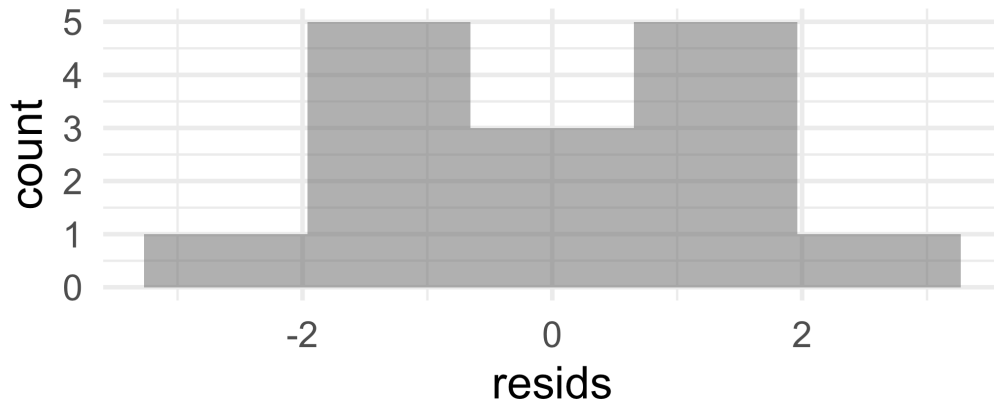
```
mod <- lm(response ~ pred1 + pred2,  
           data = my_data)
```

```
my_data <- my_data |>  
  mutate(preds = predict(mod),  
         resid = resid(mod))
```

Residuals Normal: Histogram

(Be quite generous)

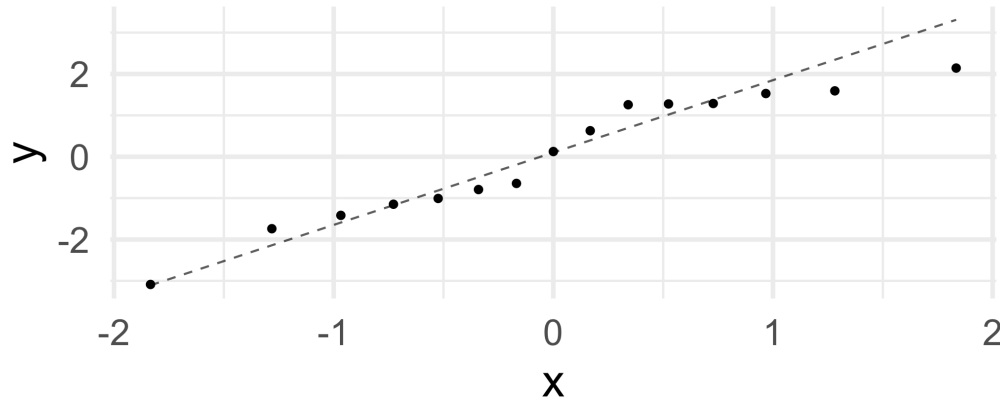
```
gf_histogram(~resids, data = my_data,  
             bins = 5)
```



Resid. Normality: Q-Q plot

(Be quite generous)

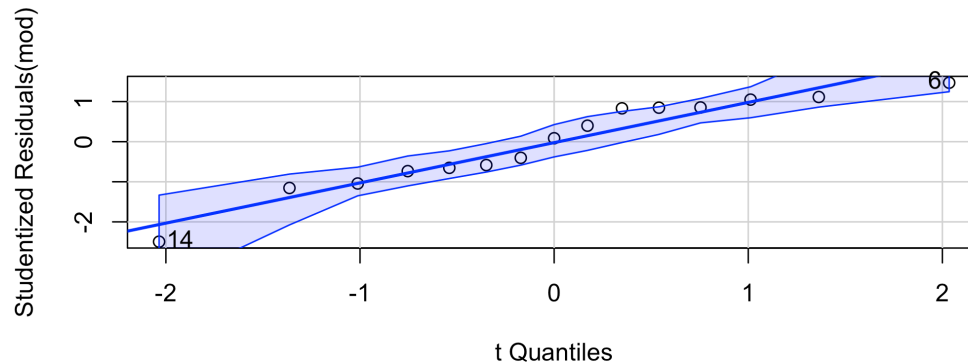
```
gf_qq(~resids, data = my_data) |>  
  gf_qqline()
```



Normality of Residuals: Q-Q plot w/CI

(Does data go far outside the CI (expected range)?)

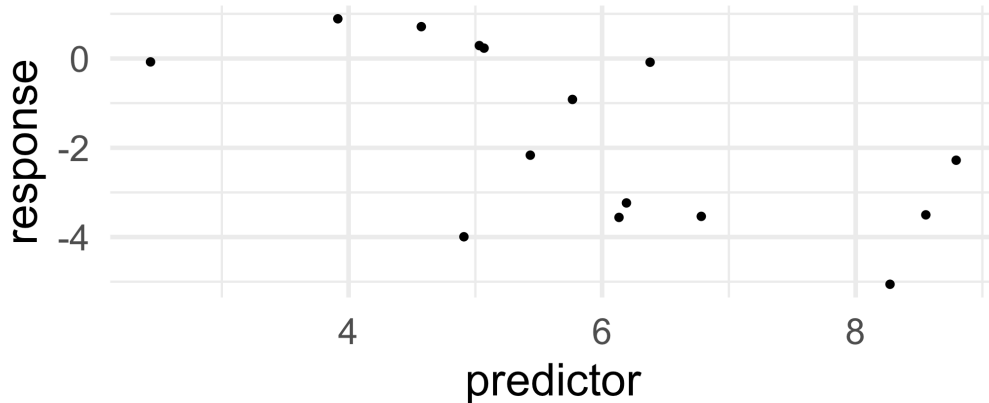
```
car::qqp(mod)
```



Lack of Non-Linearity

DATA plots: No trend, OK. Linear trend, OK

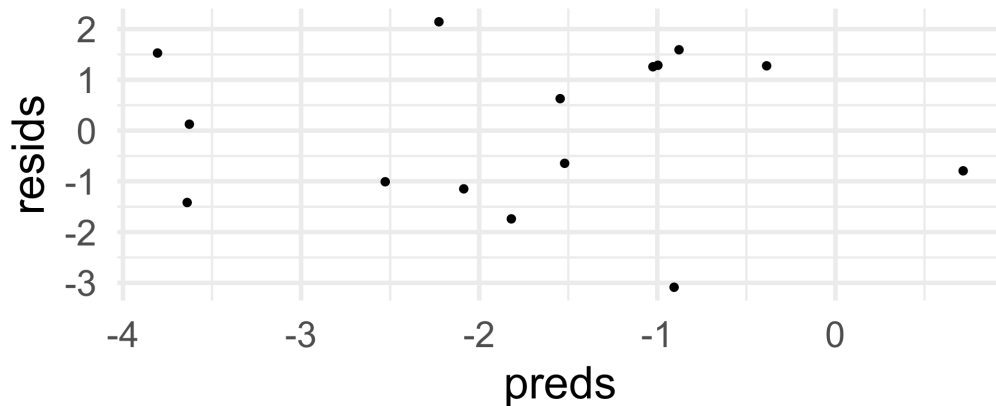
```
gf_point(response ~ predictor,  
          data = my_data)
```



Lack of Non-Linearity

RESIDUALS vs. FITTED: OK if No trends

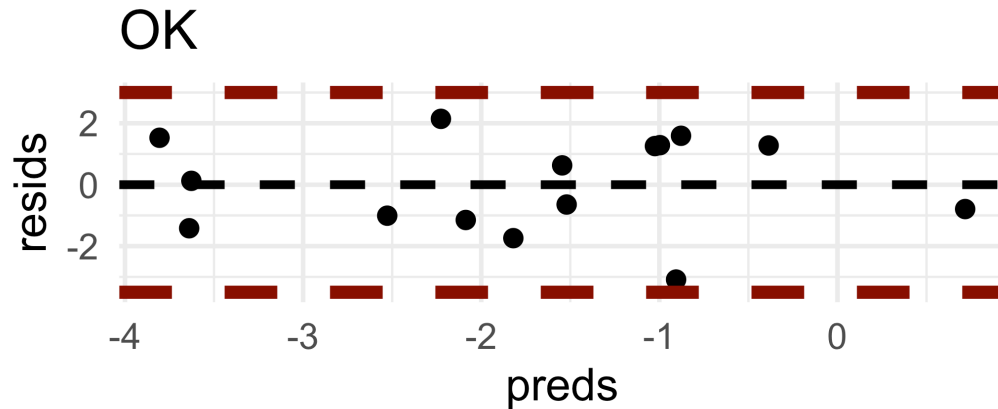
```
gf_point(resids ~ preds, data = my_data)
```



Constant Residual Variance

Point cloud should fit well in a rectangle (not trumpet)

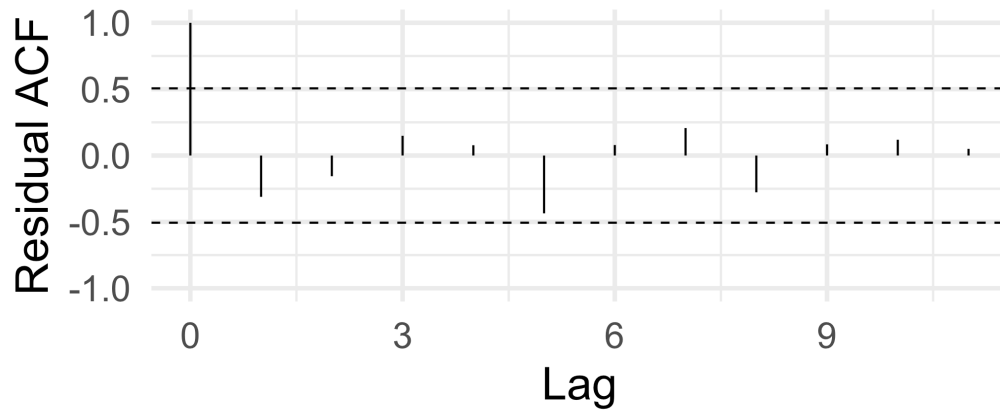
True for resid vs. predicted *and* resid vs. any predictor



Independence of Residuals

Ponder sort order; then ACF plot

```
s245::gf_acf(~mod) |>  
  gf_lims(y = c(-1,1))
```



Any LINE Violation ->

Danger!

Conclusions can not be trusted

- slope estimates **incorrect**
- CIs and p-values **too small**
- poor prediction accuracy

R-squared

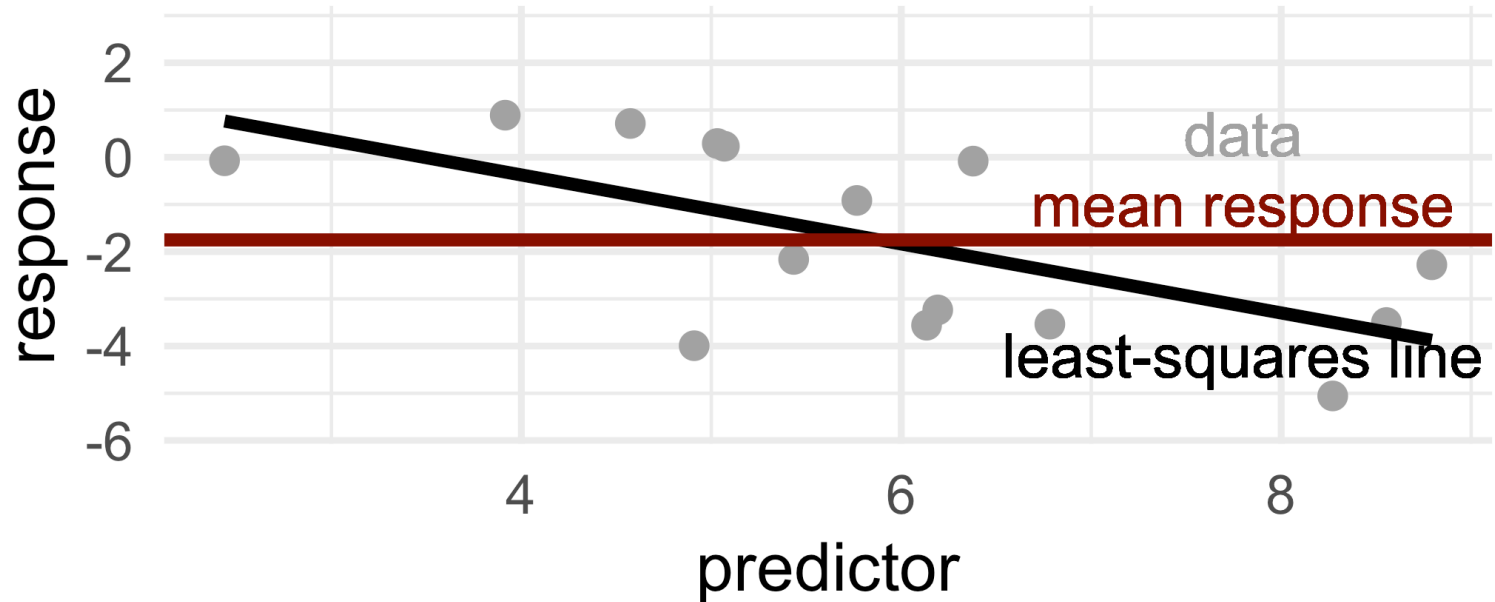
Simple measure of fit of model to data

NOT *goodness of model or appropriateness*

$$R^2 = 1 - \frac{RSS}{TSS} =$$
$$1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

R-squared

Simple measure of goodness-of-fit



R^2 ranges 0 - 1

0: no trend; 1: perfect line

Want practice? For fun, check out: <https://www.guessthecorrelation.com/>